

# 本周周报(11.26-12.2):

解聪

## 本周工作:

### 1. 时序用户行为分析

#### 通信数据的可视化探索

本周尝试对通讯数据进行可视化。数据由周霞娟提供。

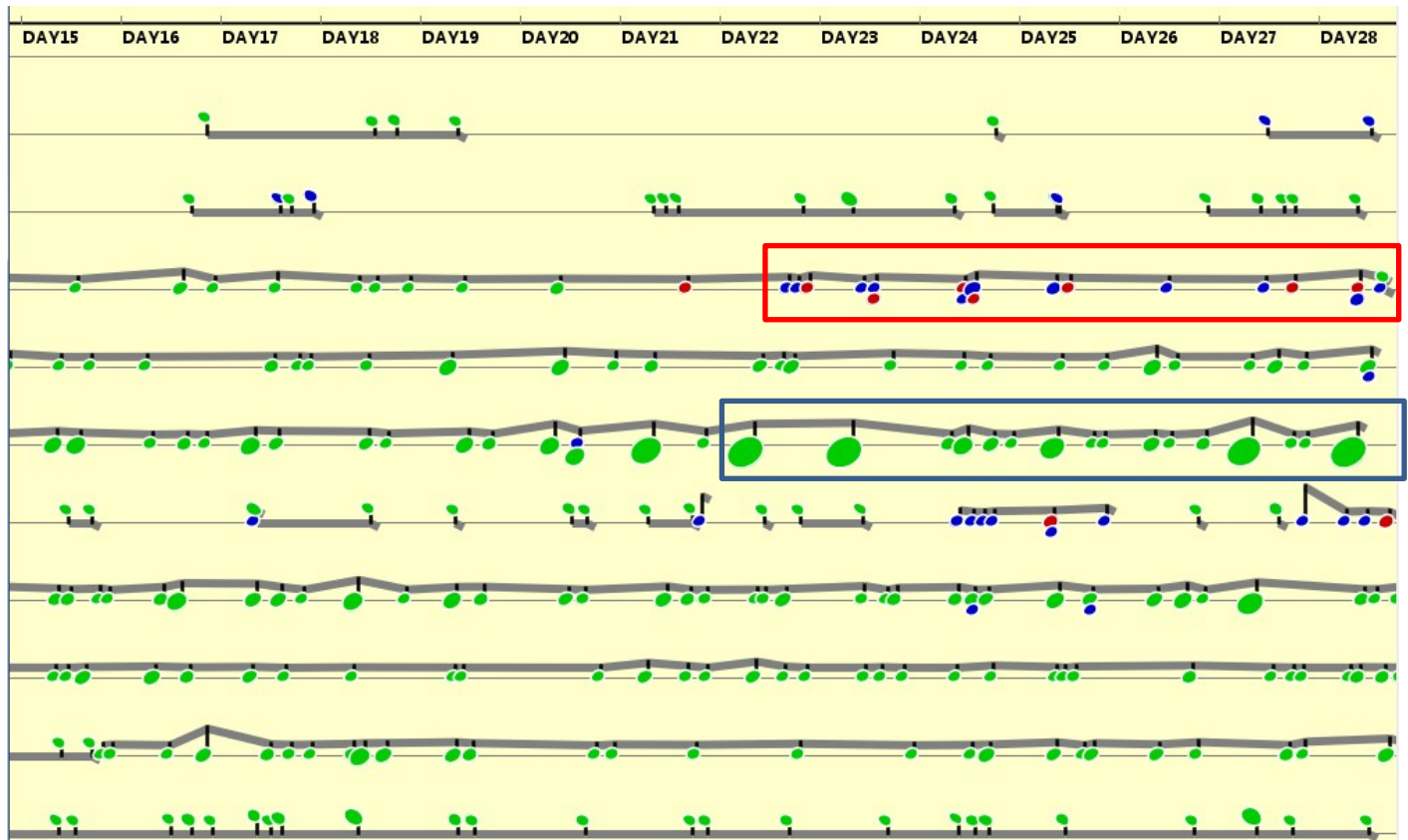
通讯数据的规模相对于淘宝的数据来说规模相对较小。目前先提取了保险行业的电话通讯记录(因为保险行业的通讯记录更可能存在一些规律)。数据的维度包括:打电话与接电话者的 ID, 通话时刻, 通话时长, 通话类型(本地, 长途等)。时序数据范围在一个月(2011 年 2 月), 共有近 2000 条的记录。

直接采用音符式可视化对数据进行分析而略去数据分析阶段。没有进行数据分析的部分原因在于: 1.数据量不是很大, 完全可以可视化所有数据; 2.问题的目标不明确, 不太清楚从通信数据中需要发现何种特定模式。

可视编码如下:

可视化	数据属性
音符	一组时间上相近的记录
音符大小	包含的通话次数
音符颜色	通话类型: 本地、长途等
音符指向朝上/朝下	打电话/接电话
符干长度	通话时长
横向位置	通话开始时间
相连音符	来自相同 ID 的记录

可视化结果如下:

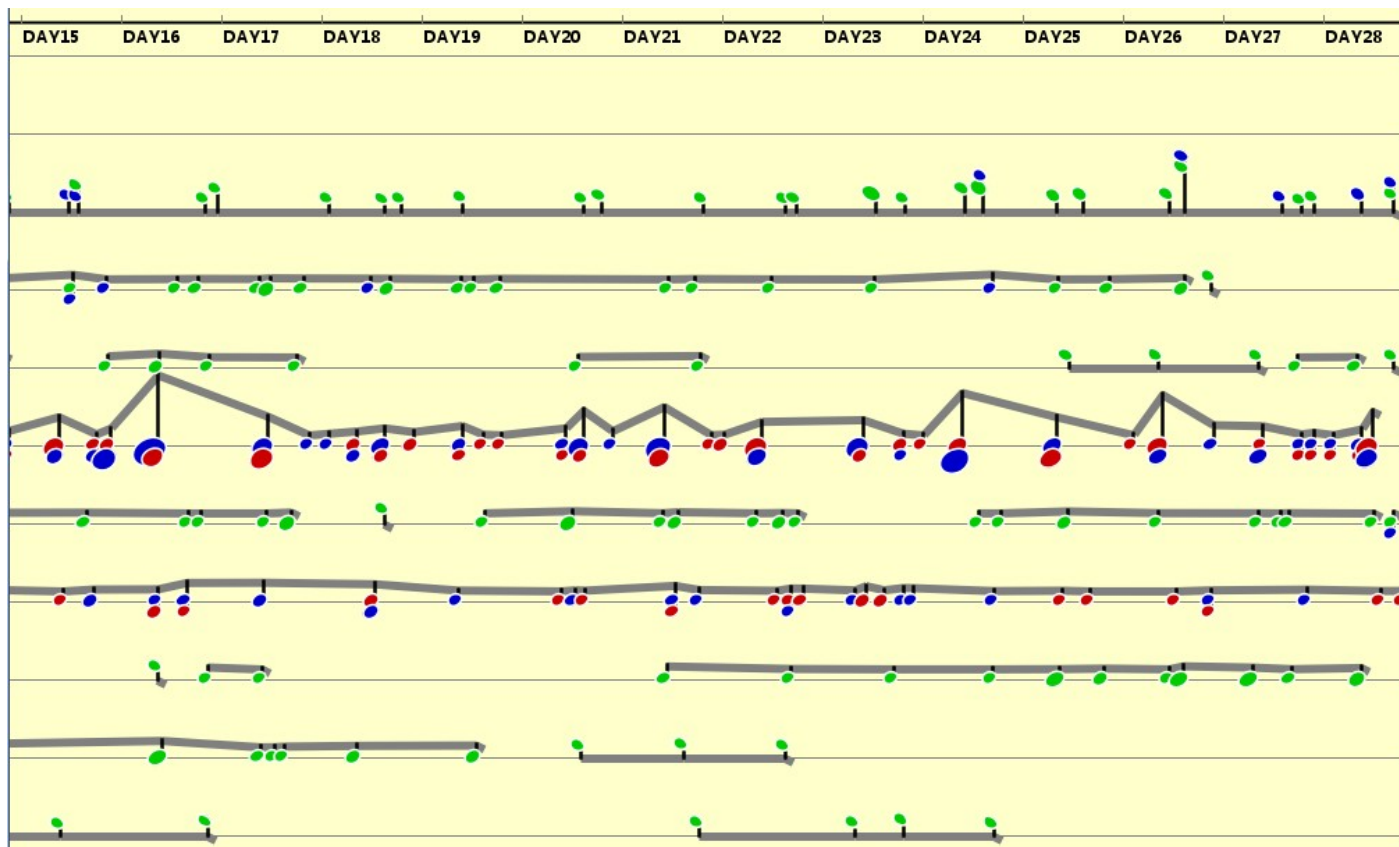


从上图可以看出：

1. 大部分用户的通话都是连续且有规律的，比如用户的通话时间与通话频率在一个月内基本不会变化太多。
2. 月末的通话存在一定的特殊规律：如蓝色方框所示月末通话频率（音符的大小）与时间（符干的长度）突然上涨；如红色方框所示月末通话类型发生变化，由本地（绿色）变为长途（蓝色）。
3. 相对于交易数据，总体来说用户通话的模式不是很明显。

另外一个很重要的问题是：可能所发现的通话行为与具体行业无太大关系。工作电话与行业的关系可能更加紧密，但这些通话可能是使用私人电话进行。根据保险行业通话的这些现有的时序数据，可以发现的信息还是有限。因此提取另外一个行业的数据作比较。

提取了 IT 行业的通话记录，以同样的方式做可视化。相对于保险行业，通话记录要少一些，同样的用户数情况下有大约 1000 条记录。

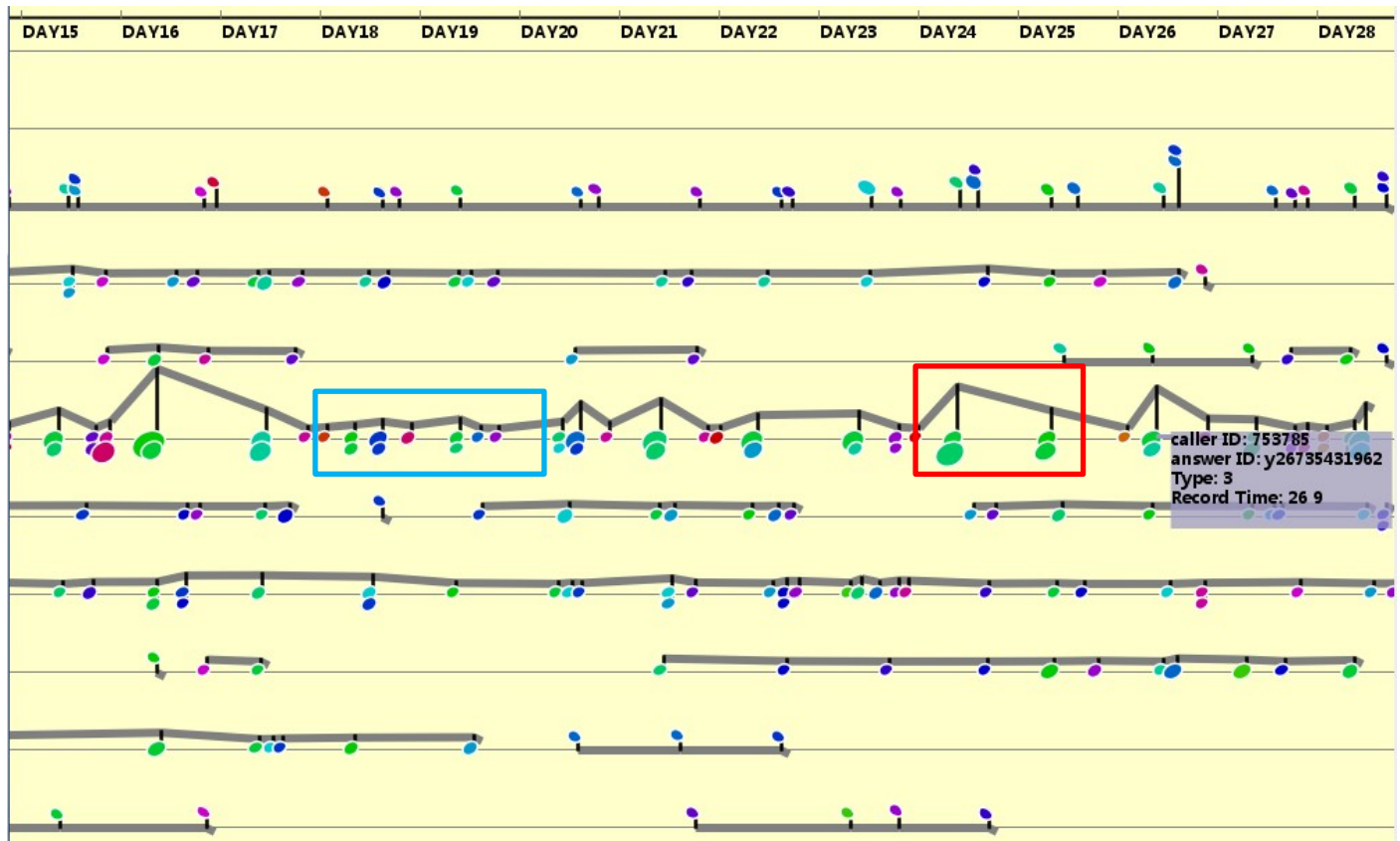


由于所获得的属性并不多，仅仅包括通话类型，通话时间等。因此只能从通话的时间信息中发现一些特征，为了验证每天通话是否有规律性，使用颜色编码每天的小时信息：



从红色到紫色的色相对应于 0 点到 23 点。绿色就对应中午午饭时间，蓝色就对应晚上下班以后的时间。

可视化结果如下：



可以看出此图与上图是同样的数据，可视化唯一的不同之处在于使用一天内的时间编码颜色。可以看出中间的一组通话记录中绿色的通话始终比较大，而且其符干较高。表示其在每天的 9-11 点左右有规律地通话，而且通话时间比平时都要长（如红色方框所示）。但是该通话记录也有例外，如蓝色方框所示，其绿色的通话音符较小。蓝色方框所对应 18,19 日是周六周日，因此可能是因为周末的关系他改变了他的通话方式。

## 通讯数据总结

1. 通讯数据与交易数据不同之处在于很少出现密集的通讯记录，更多的是用户长期的通讯行为模式。这种特征也往往需要和用户之间的通话关系结合在一起分析。
2. 通讯数据每条记录中包含的维度不多，大部分信息是有关于用户属性的。因此，可能更适合于发现用户本身的特征，或用户之间的社交网络关系。
3. 音符可视化更能够体现频率有关的信息。可能其他时序数据更适合音符的可视化。比如微博数据等，还有待验证。

## 淘宝交易分析系统的改进

添加了 calendar view，如下图所示：

September						
Mo...	Tu...	We...	Th...	Fri...	Sat...	Su...
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

因为目前拿到的数据只有一周的数据，因此只有 19-25 号有对应的颜色。代表其当天的销售额的大小。

## 下周工作：

1. 英文论文的书写。

2. 修改浙大学报工学版的投稿论文。
3. 音符可视化对其他数据的验证。
4. 组会主题报告。